TeraGrid as a Large Facility: Management Challenges *and* Blue Waters Petascale Computing Facility

John Towns
Chair, TeraGrid Forum
Director, Persistent Infrastructure
National Center for Supercomputing Applications
University of Illinois

What is this (are these) talk(s) about?

- TeraGrid as a Facility
  - Brief introduction to TeraGrid
    - what it is
    - challenges in managing the project
  - How the project is managed and how the management evolved
    - post-facto application of project management: one project's horror story
- Blue Waters Petascale Computing Facility
  - a bit about what Blue Waters is
  - some information about constructing the Petascale Computing Facility
    - can only share some information ☹

TeraGrid as a Large Facility: Management Challenges

John Towns
Chair, TeraGrid Forum
Director, Persistent Infrastructure
National Center for Supercomputing Applications
University of Illinois

TeraGrid™

Our Vision of TeraGrid

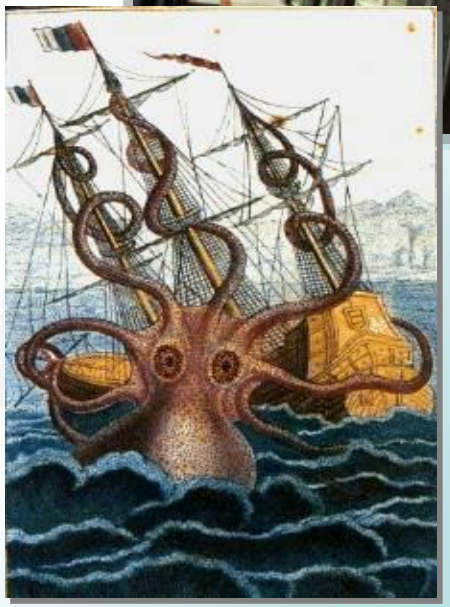
- Three part mission:
  - support the most advanced computational science in multiple domains
  - empower new communities of users
  - provide resources and services that can be extended to a broader cyberinfrastructure
- TeraGrid is…
  - an advanced, nationally distributed, open cyberinfrastructure comprised of supercomputing, storage, and visualization systems, data collections, and science gateways, integrated by software services and high bandwidth networks, coordinated through common policies and operations, and supported by computing and technology experts, that enables and supports leading-edge scientific discovery and promotes science and technology education
  - a complex collaboration of over a dozen organizations and NSF awards working together to provide collective services that go beyond what can be provided by individual institutions

TeraGrid: greater than the sum of its parts…

- Single unified allocations process
- Single point of contact for problem reporting and tracking
  - especially useful for problems between systems
- Simplified access to high end resources for science and engineering
  - single sign-on
  - coordinated software environments
  - uniform access to heterogeneous resources to solve a single scientific problem
  - simplified data movement
- Expertise in building national computing and data resources
- Leveraging extensive resources, expertise, R&D, and EOT
  - leveraging other activities at participant sites
  - learning from each other improves expertise of all TG staff
- Leadership in cyberinfrastructure development, deployment and support
  - demonstrating enablement of science not possible without the TeraGrid-coordinated human and technological resources

Diversity of Resources (not exhaustive)

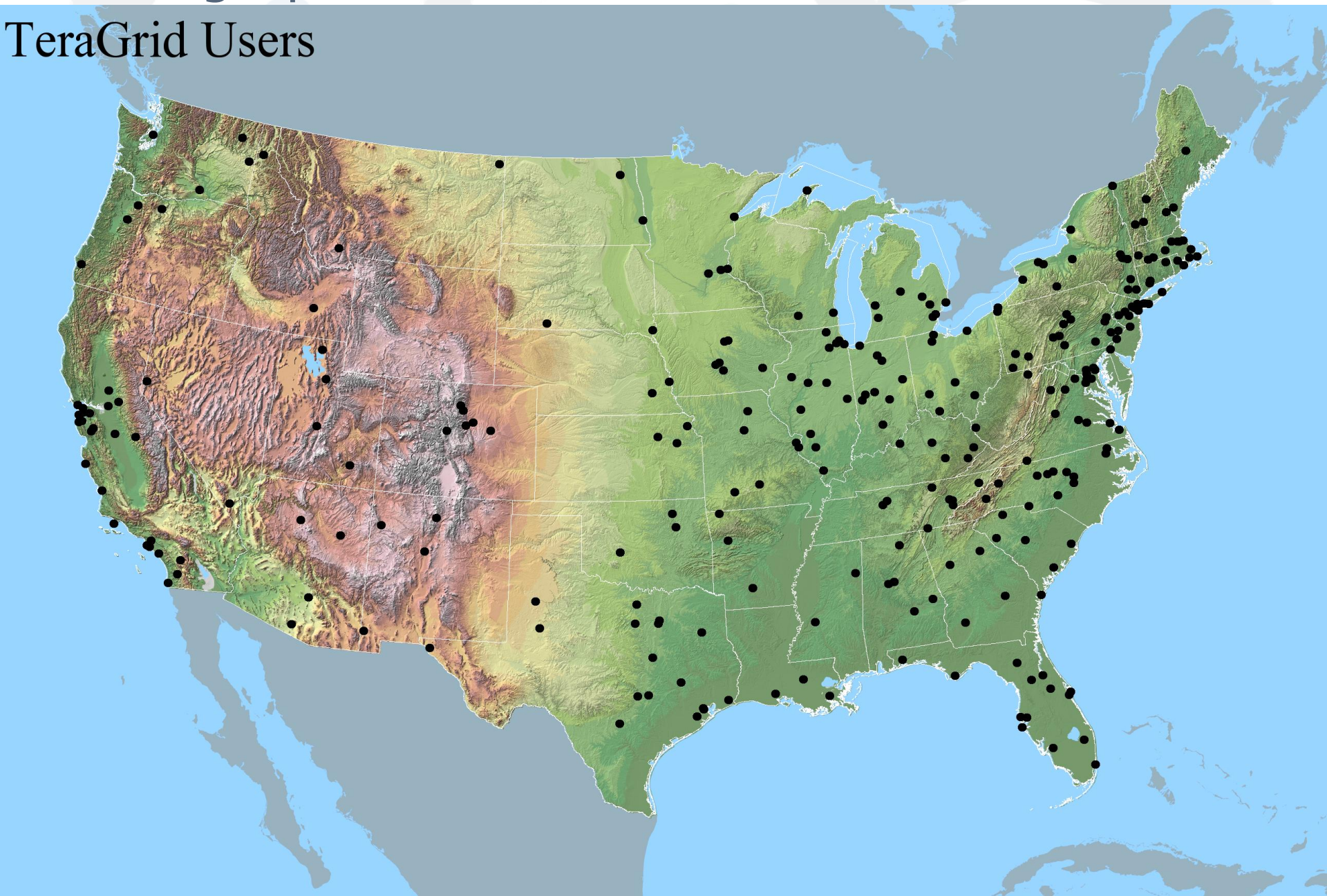
- Very Powerful Tightly Coupled Distributed Memory
  - Ranger (TACC): Sun Constellation, 62,976 cores, 579 Tflop, 123 TB RAM
  - Kraken (NICS): Cray XT5, 66,048 cores, 608 Tflop, > 1 Pflop in 2009
- Shared Memory
  - Cobalt (NCSA): Altix, 8 Tflop, 3 TB shared memory
  - Pople (PSC): Altix, 5 Tflop, 1.5 TB shared memory
- Clusters with Infiniband
  - Abe (NCSA): 90 Tflops
  - Lonestar (TACC): 61 Tflops
  - QueenBee (LONI): 51 Tflops
- Condor Pool (Loosely Coupled)
  - Purdue- up to 22,000 cpus
- Visualization Resources
  - TeraDRE (Purdue): 48 node nVIDIA GPUs
  - Spur (TACC): 32 nVIDIA GPUs
- Storage Resources
  - GPFS-WAN (SDSC)
  - Lustre-WAN (IU)
  - Various archival resources



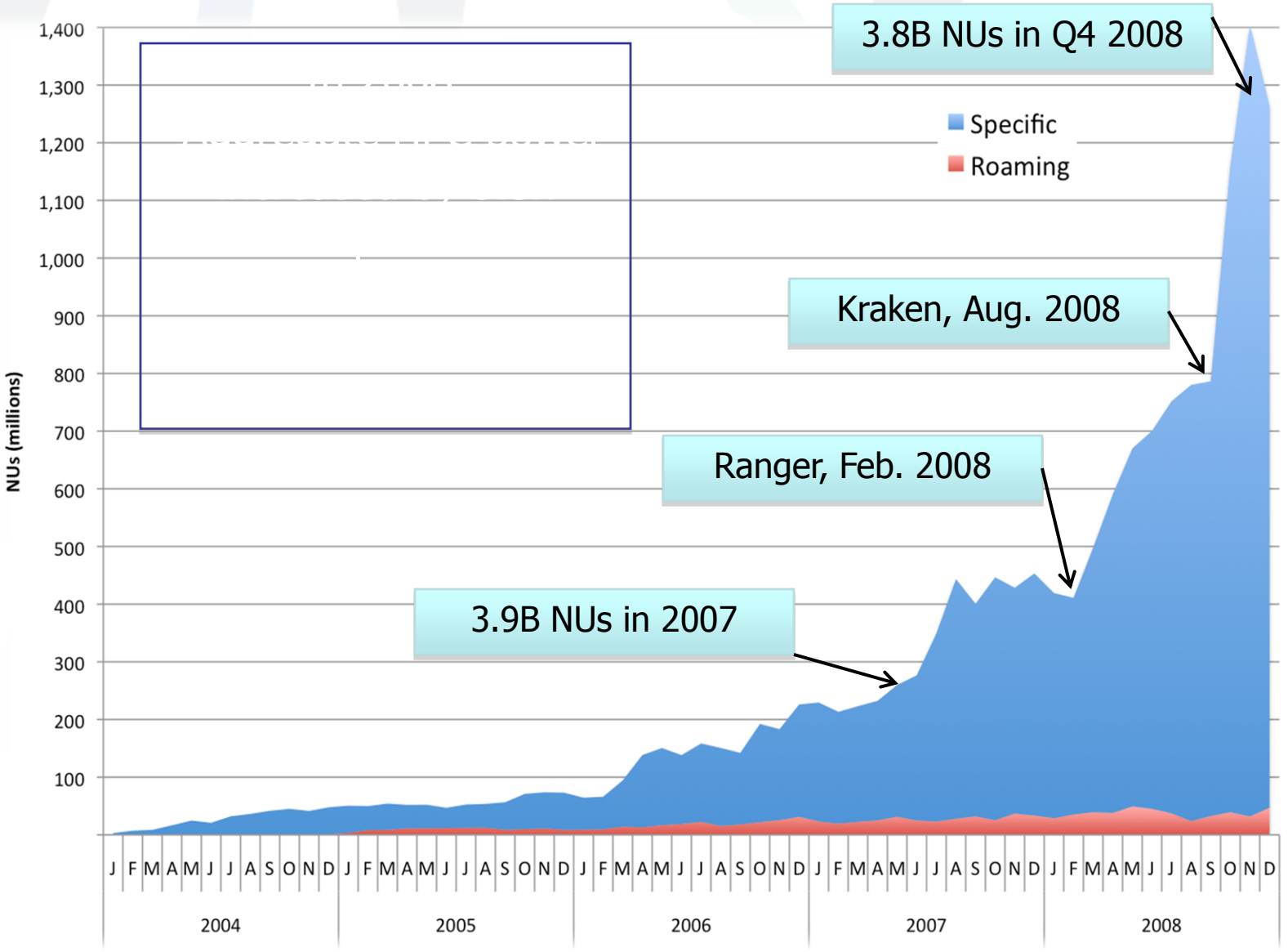
Resources to come…

- Track 2c @ PSC
  - large shared memory system in 2010
- Track 2d being competed
  - data-intensive HPC system
  - experimental HPC system
  - pool of loosely coupled, high throughput resources
  - experimental, high-performance grid test bed
- eXtreme Digital (XD) High-Performance Remote Visualization and Data Analysis Services
  - service and possibly resources; up to 2 awards (?)
- Blue Waters (Track 1) @ NCSA:
  - 1 Pflop sustained on serious applications in 2011
- Unsolicited proposal for archival storage enhancements pending

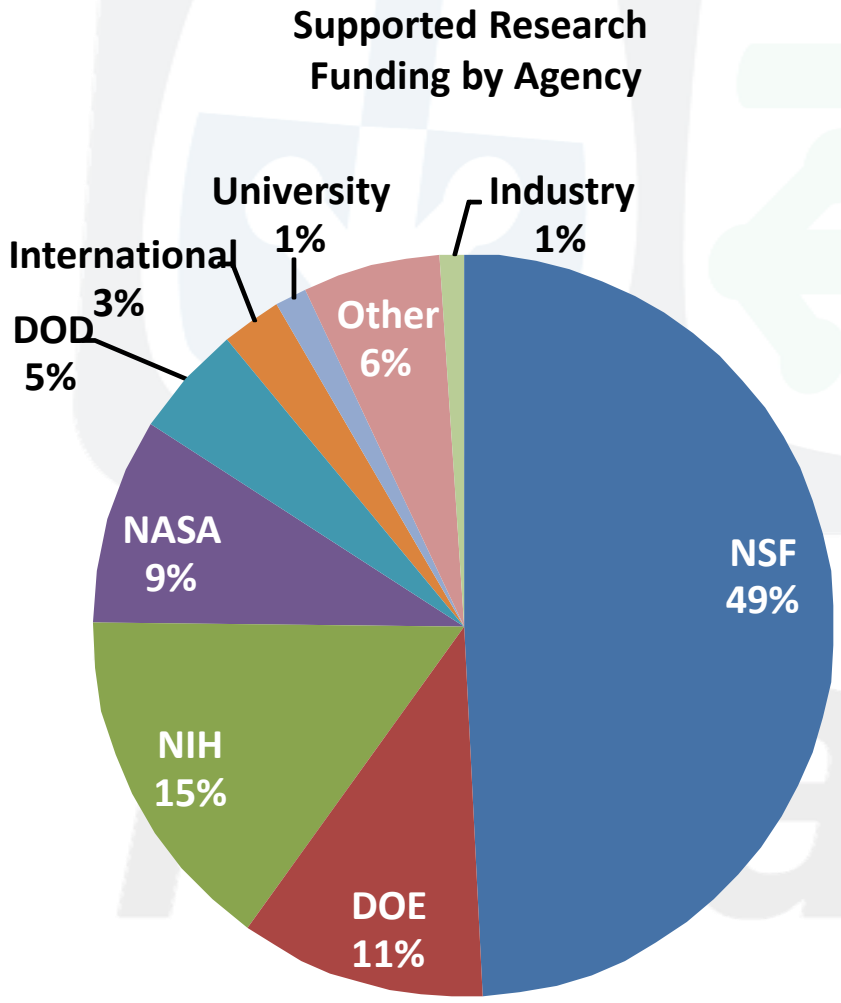
Geographical Distribution of TeraGrid Users

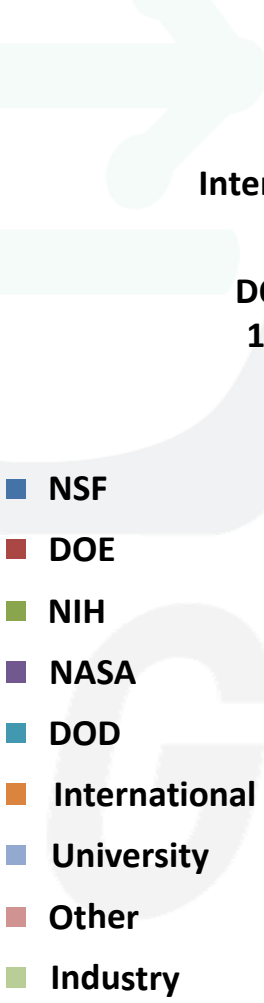


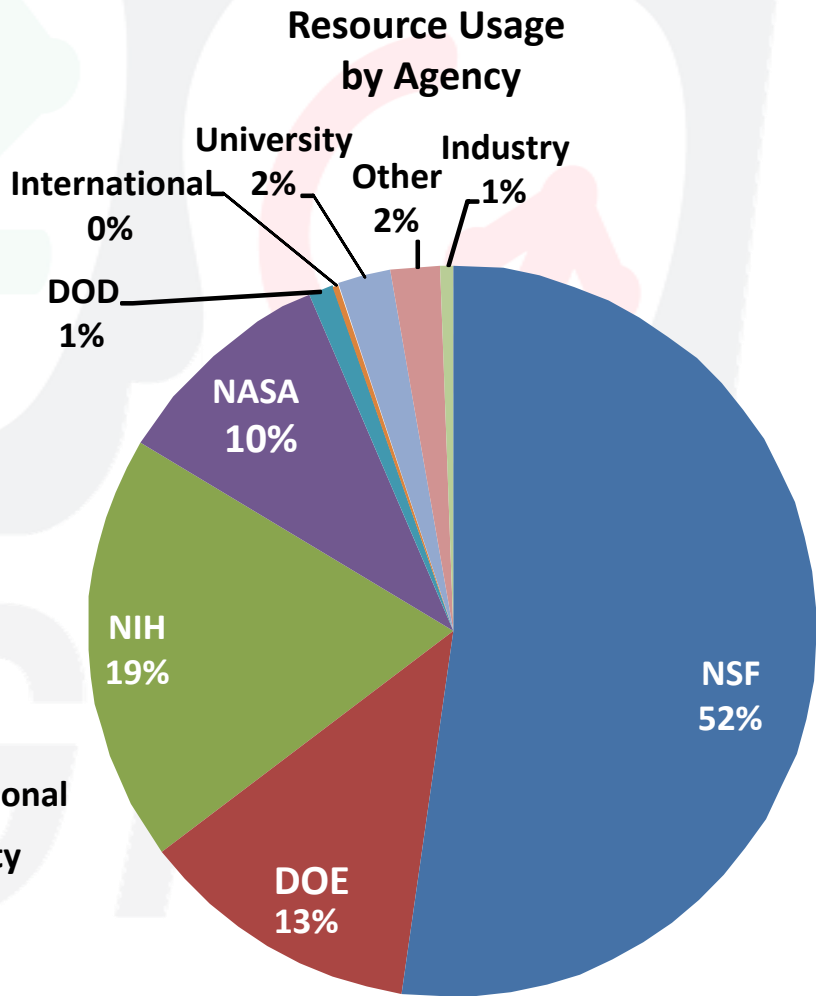
TeraGrid HPC Usage, 2008



Impacting Many Agencies







$91.5M in Funded Research Supported

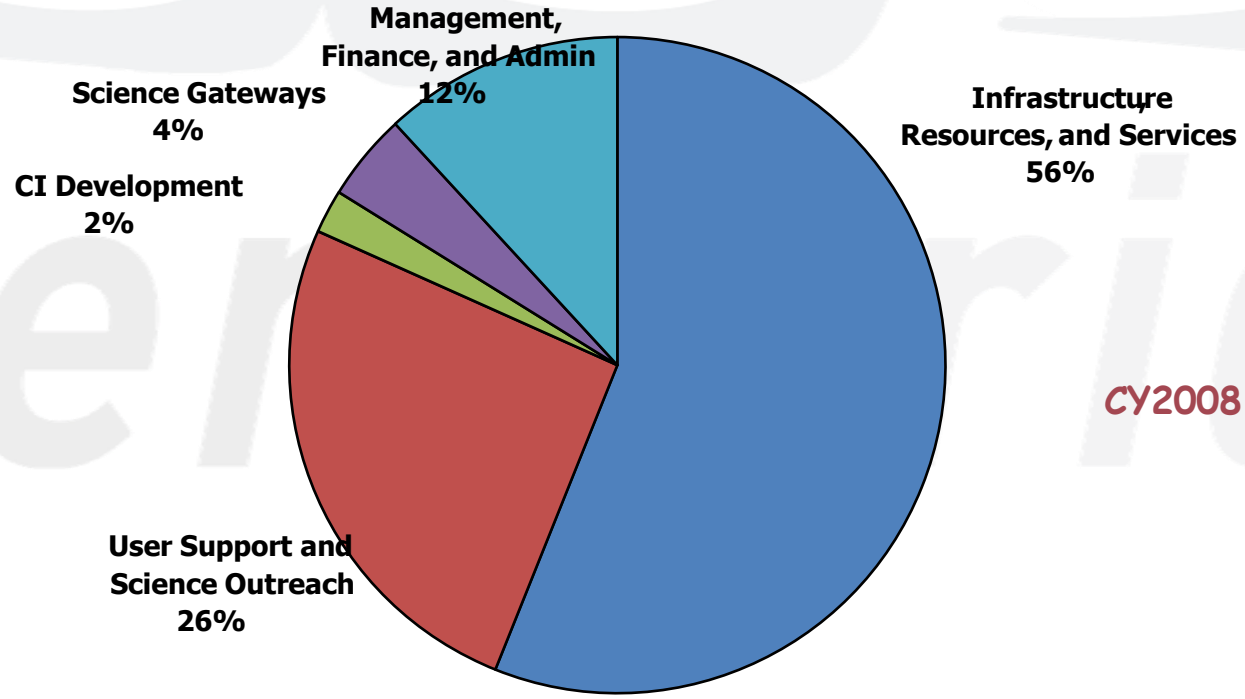
10B NUs Delivered

Special Challenges in TeraGrid

- Federated project
  - 12 awards, 13 institutions, >225 FTEs
  - one project
- Long range planning is VERY difficult
  - intent is to operate bleeding edge resources and services
  - often introduced as new NSF awards through NSF review process
    - we usually know something is coming, but we don't know what it is and who is bringing it
- Mis-match between project and imposed project management practice
  - little to no project management at outset of project!!
  - traditional project management suited to design-bid-build construction project
  - TeraGrid is a combination of operational activities and short-term integration/development projects

CY08 Total TeraGrid Expenditures Distribution

- Distribution of total TeraGrid expenditures closely resembles RP expenditures; RP expenditures ~4x those of the GIG
- The bulk of TeraGrid funds go directly to providing facilities to scientific researchers and making sure that they have the support needed so that they can make productive use of them

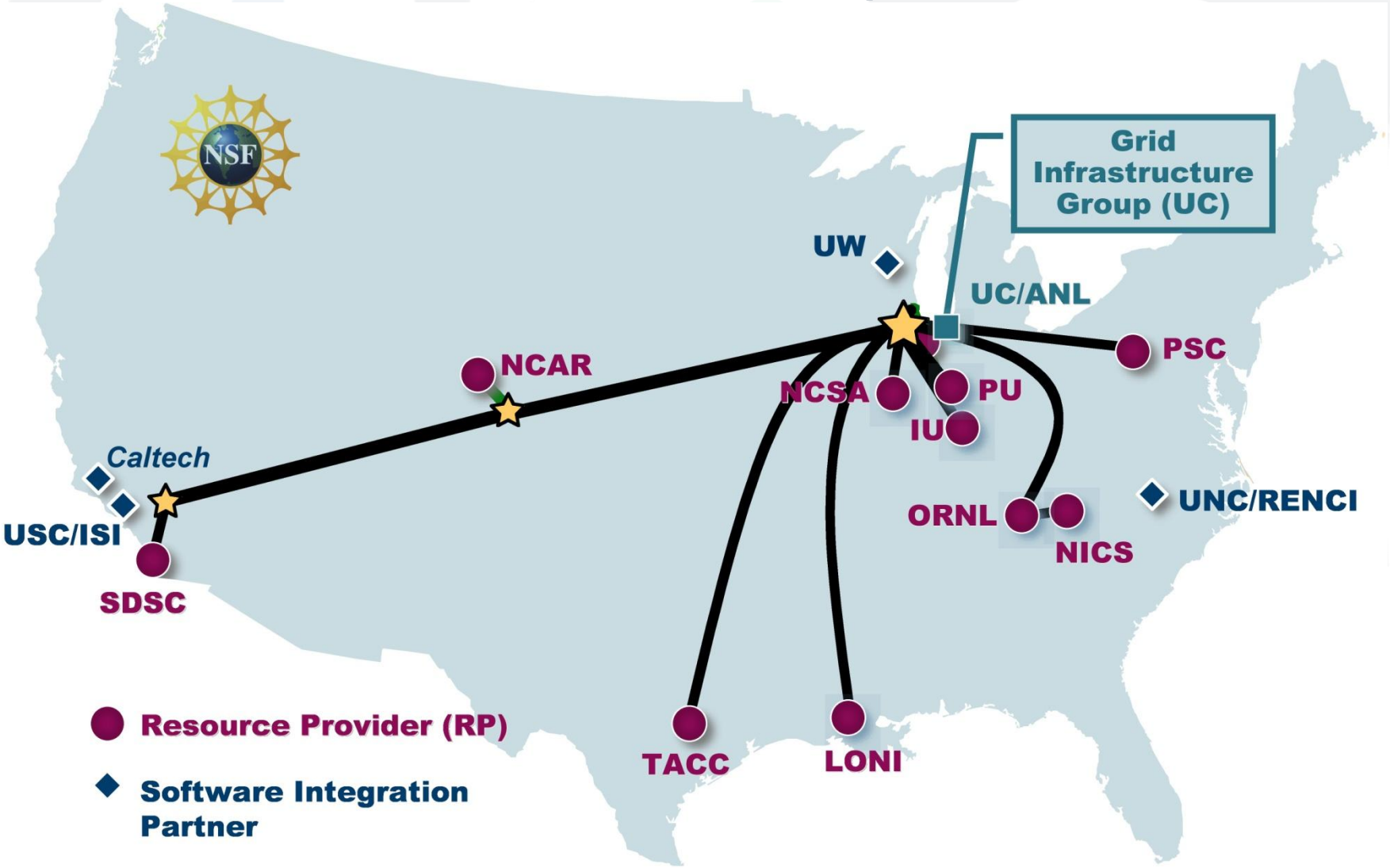


CY2008: $50.2M

How is TeraGrid Organized?

- TG is set up like a large cooperative research group
  - evolved from many years of collaborative arrangements between the centers
  - still evolving!
- Federation of 12 awards
  - Resource Providers (RPs)
  - Grid Infrastructure Group (GIG)
- Strategically lead by the TeraGrid Forum
  - made up of the PI's from each RP and the GIG
  - led by the TG Forum Chair, who is responsible for coordinating the group (elected position)
    - John Towns – TG Forum Chair
  - responsible for the strategic decision making that affects the collaboration
- Centrally coordinated by the GIG

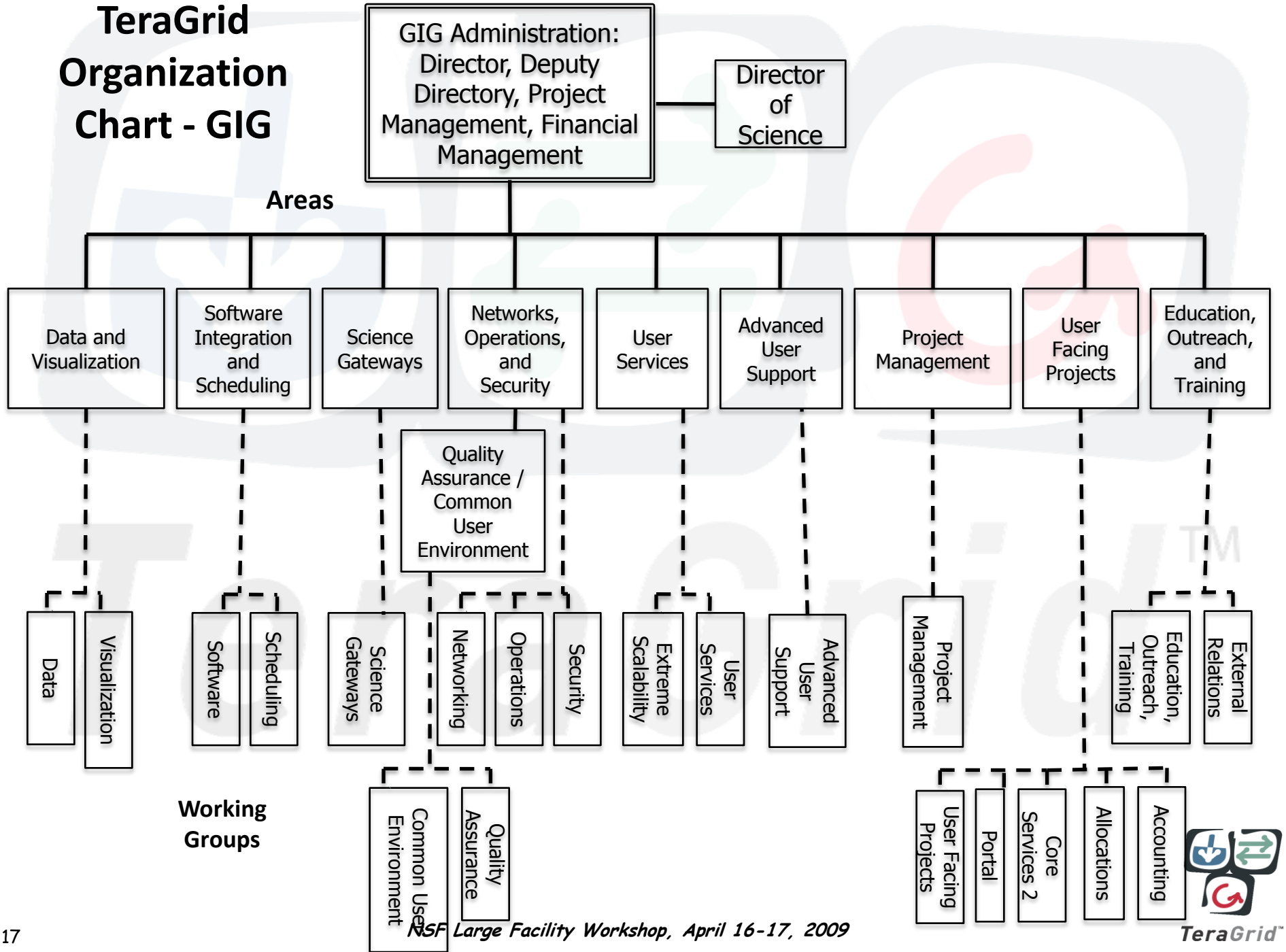
TeraGrid Participants



Who are the Players?

- GIG Management
  - GIG Director: Matthew Heinzel
  - GIG Director of Science: Dan Katz
  - Area Directors:
    - Software Integration: Lee Liming/J.P. Navarro
    - Gateways: Nancy Wilkins-Diehr
    - User Services: Sergiu Sanielevici
    - Advanced User Support: Amit Majumdar
    - Data and Visualization: Kelly Gaither
    - Network, Ops, and Security: Von Welch
    - EOT: Scott Lathrop
    - Project Management: Tim Cockerill
    - User Facing Projects and Core Services: Dave Hart
- TeraGrid Forum
  - TG Forum Chair: John Towns
  - Membership:
    - PSC: Ralph Roskies
    - NICS: Phil Andrews
    - ORNL: John Cobb
    - Indiana: Craig Stewart
    - Purdue: Carol Song
    - U Chicago/ANL: Mike Papka
    - NCSA: John Towns
    - LONI: Dan Katz
    - TACC: Jay Boisseau
    - NCAR: Rich Loft
    - SDSC: Richard Moore
    - GIG: Matt Heinzel

TeraGrid Organization Chart - GIG



Communications are Key

- TeraGrid Round Table Meeting
  - bi-weekly for members of all TeraGrid groups via Access Grid
  - rotating presentations
    - Working Groups, Resource Providers, etc.
    - information of general interest
- GIG Area Director Calls
  - weekly call for GIG administration and Area Directors
  - coordinates GIG activities
- TG Forum Call
  - Bi-weekly call for RP PIs or site leads with GIG ADs invited
  - coordinates across GIG and RPs
- Working Groups
  - periodic calls conducted by each Working Group to coordinate activities
  - supports regular communication via e-mail and TG wiki
- TeraGrid Quarterly Management Meetings
  - two day meetings in March, June, September, and December
    - GIG and TG Forum staff; any other interested TG stakeholders
  - coordinates activities across GIG and RPs, including annual and quarterly planning and reporting
- Science Advisory Board Meeting
  - two day meeting each January and June of joint TeraGrid / NSF Science Advisory Board
  - reviews TeraGrid accomplishments and plans to obtain feedback/guidance from the scientific community
- Quarterly Allocation Meetings
  - two day meetings in March, June, September, and December for TRAC peer review panel and sire representatives
  - reviews and decides on TeraGrid resource allocation proposals received through allocations requests
- TeraGrid Annual Conference (TGxy)
  - four day conference each June open to all TeraGrid participants and the scientific and educational communities
  - showcases the capabilities, achievements, and impact of TeraGrid in research and education through presented papers, demonstrations, posters, and visualizations

Evolution of Project Management

- Initially, GIG and each RP had separate PM activities which were loosely coupled
  - separate NSF research awards and project management
- Separate planning functions and reports, until PY3 started integrated annual reporting/review
  - TeraGrid starts adapting traditional project management methods so that they apply to these research awards
- GIG mainly project-based, RPs operations-based
  - Traditional project management methods suitable for projects – deliverables and end dates
  - Operations are ongoing support activities
- PY4 IPP (Integrated Project Plan)
  - result of April 2008 review
  - WBS too detailed to see forest for the trees
  - Program Plan too high level
  - amalgamation of GIG plans and RP plans
- Established TeraGrid Project Management Area
  - assigned new Area Director
  - formed Project Management Working Group – more tightly coupled project management
- Integrated GIG/RP planning process developed as evolution of GIG formalized project planning process
  - PY5 IPP is first plan developed utilizing the fully integrated GIG/RP planning process
- Now have more tightly coupled, fully integrated project management processes for
  - planning
  - tracking
  - reporting
  - managing change

TeraGrid Integrated Planning: Strategic Objectives

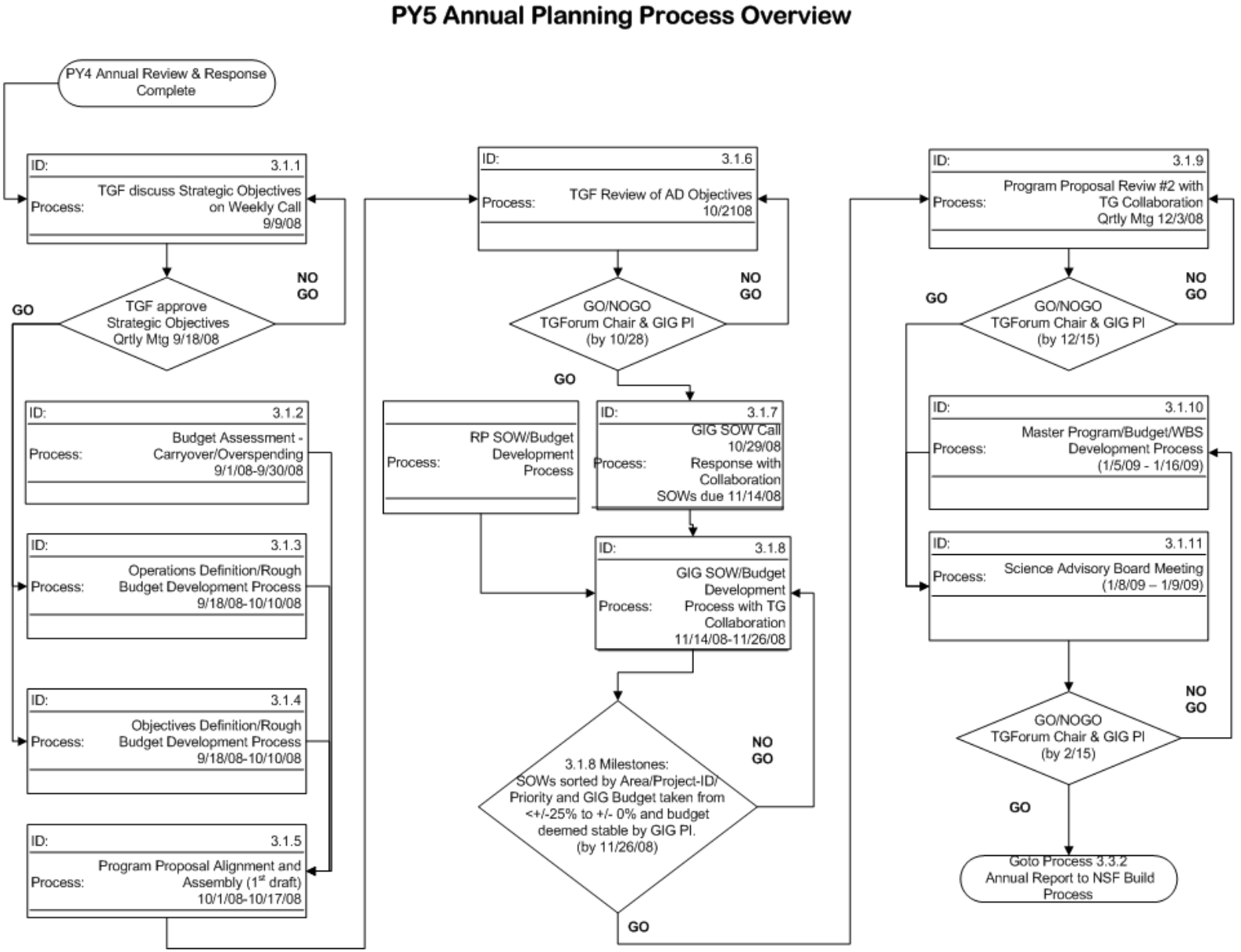
- Objectives determined from considering numerous inputs
  - user input via various mechanisms
    - surveys, user contacts, advisory bodies, review panels, etc.
  - technical input from TG staff

- Planning for PY5 started by identifying 5 high level project strategic objectives (no change from PY4)
  - Enable science that could not be done without TeraGrid
  - Broaden the user base
  - Simplify users lives
  - Improve Operations
  - Enable connections to external resources

TeraGrid RP/GIG Integrated Planning Process

- TG Forum communicates strategic objectives
- Develop operational and project objectives
  - high-level objectives to accomplish before the end of the project
    - each objective tied back to at least one strategic objective
- Statements of Work
  - call for SOW's made
    - scope and deliverables
    - FTE effort and staff names
    - cost
- Budget created
  - based on SOWs
    - budget ties directly to SOWs
    - each SOW item addresses one or more AD objectives
    - each objective ties to one or more strategic objectives
- All the pieces combined into a detailed Integrated Project Plan (IPP)
  - integrate RP and GIG plans
    - identify synergistic RP/GIG activities
    - identify duplicate efforts
    - build integrated Work Breakdown Structure and Budget
  - final review by TeraGrid Forum
  - review by Science Advisory Board
  - Project managers develop final version of IPP
- Annual Program Plan based on IPP

Planning Flow Chart



Reporting

- Integrated Program Plan (IPP)
  - primary output of the yearly TeraGrid planning process
  - documents the objectives, budget, and activities for the next project year for the GIG and the RPs
- Work Breakdown Structure (WBS)
  - part of the IPP that lists milestones and resources for all planned TeraGrid activities
- Quarterly Reports
  - report by Area of TeraGrid activities during the previous quarter (1, 2, and 3) of the calendar year
- Annual Report and Program Plan
  - comprehensive report of TeraGrid activities during the previous calendar year and plan for the next project year
- Monthly Resource Provider System Statistics reports

Blue Waters and the Petascale Computing Facility

National Center for Supercomputing Applications
University of Illinois at Urbana Champaign

Criteria for Petascale Computing System

- Maximize Core Performance
  - … to minimize the number of cores needed for a given level of performance as well as the impact of sections of code with limited scalability
- Incorporate Large, High-bandwidth Memory Subsystem
  - … to enable the solution of memory-intensive problems
- Optimize Interconnect Performance
  - … to facilitate scaling to the large numbers of processors required for sustained petascale performance
- Integrate High-performance I/O Subsystem
  - … to enable solution of data-intensive problems
- Maximize System Integration, Leverage Mainframe Reliability, Availability, Serviceability (RAS) Technologies
  - … to assure reliable operation for long-running, large-scale simulations

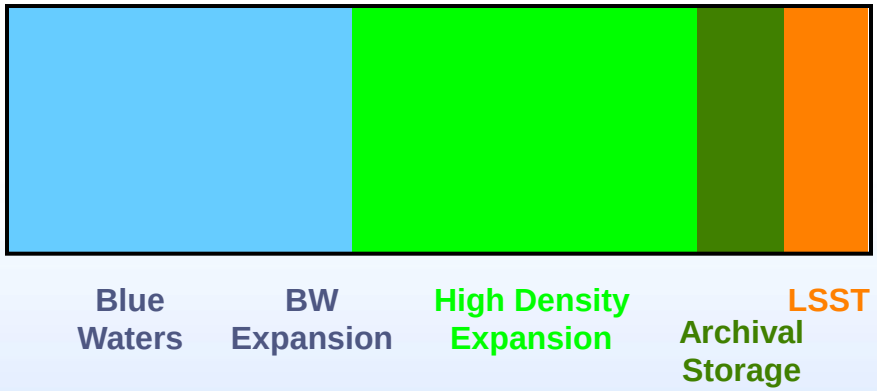
Challenges in Petascale Computing

- Hardware Issues
  - Scale of petascale system ~*100x* existing system
    - "Quantity has a quality all its own"
  - Petascale systems requires most advanced computing technologies
    - Microprocessor and memory (high performance, low latency, high bandwidth)
    - Interconnect (low latency, high bandwidth)
    - I/O subsystem (high bandwidth)
- Software Issues
  - Computing systems software
    - System software (scalability, control of jitter, *etc*.)
    - Software development environment and tools
    - Reliability, availability, and usability
  - Analysis and visualization of tera-petascale datasets
  - Scientific and engineering models and applications
    - Few algorithms are scalable to 100,000s of processors
    - Need high *sustained* performance on demanding S&E applications

Blue Waters Petascale Computing System

- Blue Waters
  - Based on IBM PERCS
  - 1 petaflops *sustained* performance
  - Multicore chips, >200,000 cores
  - >800 terabytes of memory
  - >10 petabytes of user disk storage
  - Water cooled
  - < 5,000 ft^2
  - On-line: July 2011

Machine Room Layout



Petascale Computing Facility
Short Description

- Design-Bid-Build construction project
- State-of-the-art, Greenfield data center designed for Blue Waters
- The project is on schedule and budget
- Substantial completion expected May 4, 2010

PCF Team Members

Gensler

CLAYCO
THE ART & SCIENCE OF BUILDING

- UI: project management; engineering and architectural standards/review; utilities design and construction
- EYPMCF/Gensler: Architect and Engineer
- Clayco/Nova: Construction Manager
- IBM: provide system specifications, third party consulting services and extensive facilities expertise
- NCSA: provide user requirements; manage the managers; focus on schedule and budget

Illinois Petascale Computing Facility

PCF Overview

- $72.5M project budget
- 93,056 GSF over two stories—45’ tall
- 30,000+ GSF of raised floor
- Office space for up to 50
- 24MW electrical capacity
- 5,400 tons cooling capacity
- Three on-site cooling towers
- Slots for 48 CRAH units
- USGBC LEED Silver classification target (Native+5)
- Began demolition, earthwork in October
- Begin deep foundation work in early November
- Phase two contracts out in December
- Five acre site allows room for facility expansion

PCF Challenges

- East Central Illinois is a tornado zone
  - many tornadoes and severe storms every spring
  - structure must be resilient to effects
  - PCF can withstand an F3 tornado—165 mph wind
- Physical security a greater concern that for previous data centers (at academic sites)
  - substantial security: cameras, digital video recording, biometrics, perimeter
  - PCF physical security equivalent to FBI field office
- Green design a big concern
  - LEED Silver classification target
    - prairie restoration area surrounding the building
    - DC power direct to compute racks: ~15% power savings!
    - water cooled system: ~40% energy use reduction!
    - “free cooling” approximately 7 months of the year
- Will need lots of power
  - 24MW initial power to building
    - 3 x 8MW feeds
  - ability to bring in additional power at later date